Stochastic Policy Gradient Methods: Improved Sample Complexity for Fisher-non-degenerate Policies

ETH zürich

Ilyas Fatkhullin   Anas Barakat   Anastasia Kireeva   Niao He

ICML International Conference On Machine Learning

Problem Setting

Markov Decision Process (MDP):

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma).$$

Parameterized policy: π_θ , $\theta \in \mathbb{R}^d$.

$$\textbf{Goal:} \quad \max_\theta J(\theta) := \mathbb{E}_{\rho, \pi_\theta}\left[\sum_{h=0}^{+\infty} \gamma^h r(s_h, a_h)\right].$$

Assumptions

Assumption 1. Fisher-non-degenerate (FND) policy. There exists $\mu_F > 0$ such that for all $\theta \in \mathbb{R}^d$,

$$F_\rho(\theta) \succcurlyeq \mu_F \cdot I_d, \quad \text{where}$$

$$F_\rho(\theta) := \mathbb{E}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)}[\nabla \log \pi_\theta(a|s) \nabla \log \pi_\theta(a|s)^\top].$$

Examples: Gaussian and **Cauchy** policies ($\mathcal{S}$ and $\mathcal{A}$ can be continuous!)

$$\pi_\theta(a|s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(a - \varphi(s)^\top \theta)^2}{2\sigma^2}\right),$$

$$\pi_\theta(a|s) = \frac{1}{\pi\sigma}\left(1 + \left(\frac{a - \varphi(s)^\top\theta}{\sigma}\right)^2\right)^{-1}.$$

Assumption 2. Compatible function approximation framework [1, 2, 3]. For all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}[(A^{\pi_\theta}(s,a) - (1-\gamma) w^*(\theta)^\top \nabla \log \pi_\theta(a|s))^2] \le \varepsilon_{\text{bias}},$$

where A^{π_θ} is the advantage function, $w^*(\theta) := F_\rho(\theta)^\dagger \nabla J(\theta)$, π^* is an optimal policy, and $\mathbb{E} \equiv \mathbb{E}_{s \sim d_\rho^{\pi^*}, a \sim \pi^*(\cdot|s)}$.

Prior Work

Sample complexities for achieving $\mathbb{E}[J^* - J(\theta)] \le \varepsilon$.

Discrete/finite $\mathcal{S}$, $\mathcal{A}$ spaces	FND policies (Assumptions 1 and 2)
TSIVR-PG (Q)-NPG Policy Mirror Descent	Vanilla-PG [1] SRVR-PG and Natural-PG [2] STORM-PG-F [3]
$\tilde{\mathcal{O}}(\varepsilon^{-2})$	$\tilde{\mathcal{O}}(\varepsilon^{-3})$

Question

Can we **improve** $\tilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity for FND policies using **computationally efficient** PG algorithm?

Algorithms

Algorithm 1 N-PG-IGT
Normalized-PG-with Implicit Gradient Transport [4]

1. **Input**: $\theta_0, \theta_1, d_0, T, H, \{\eta_t\}_{t\ge0}, \{\gamma_t\}_{t\ge0}$
2. **for** $t = 1, \ldots, T-1$ **do**
3. $\tilde{\theta}_t = \theta_t + \frac{1-\eta_t}{\eta_t}(\theta_t - \theta_{t-1})$
4. Sample a trajectory $\tilde{\tau}_t \sim p(\cdot|\pi_{\tilde{\theta}_t})$ of length H
5. $d_t = (1-\eta_t) d_{t-1} + \eta_t \widetilde{\nabla} J(\tilde{\tau}_t, \tilde{\theta}_t)$
6. $\theta_{t+1} = \theta_t + \gamma_t \frac{d_t}{\|d_t\|}$
7. **end for**

Algorithm 2 (N)-HARPG
(Normalized)-Hessian-Aided Recursive PG

1. **Input**: $\theta_0, \theta_1, d_0, T, H, \{\eta_t\}_{t\ge0}, \{\gamma_t\}_{t\ge0}$
2. **for** $t = 1, \ldots, T-1$ **do**
3. $q_t \sim \mathcal{U}([0,1])$
4. $\hat{\theta}_t = q_t\theta_t + (1-q_t)\theta_{t-1}$
5. Sample $\tau_t \sim p(\cdot|\pi_{\theta_t})$; $\hat{\tau}_t \sim p(\cdot|\pi_{\hat{\theta}_t})$ of length H
6. $v_t = \widetilde{\nabla}^2 J(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})$
7. $d_t = (1-\eta_t)(d_{t-1} + v_t) + \eta_t \widetilde{\nabla} J(\tau_t, \theta_t)$
8. $\theta_{t+1} = \begin{cases} \theta_t + \gamma_t d_t & \text{(HARPG)} \\ \theta_t + \gamma_t \frac{d_t}{\|d_t\|} & \text{(N-HARPG) [5]} \end{cases}$
9. **end for**

Advantages:

- Easy to implement
- No IS weights
- Single loop
- Batch-free
- Comp. efficient
- Low memory

Global Convergence

N-PG-IGT

Theorem 1. Under Assumptions 1, 2 and regularity of π_θ, if we set $\gamma_t = \mathcal{O}\left(\frac{1}{t}\right)$, $\eta_t = \frac{1}{(t+1)^{4/5}}$ and $H = (1-\gamma)^{-1}\log(T+1)$, then

$$J^* - \mathbb{E}[J(\theta_T)] \le \mathcal{O}\left(\frac{1}{(T+1)^{2/5}}\right) + \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}.$$

HARPG and N-HARPG

Theorem 2. Under Assumptions 1, 2 and regularity of π_θ, if we set $\gamma_t = \mathcal{O}\left(\frac{1}{t^{1/2}}\right)$ ($\gamma_t = \mathcal{O}\left(\frac{1}{t}\right)$ for N-HARPG), and $\eta_t = \frac{1}{t+1}$, $H = (1-\gamma)^{-1}\log(T+1)$, then

$$J^* - \mathbb{E}[J(\theta_T)] \le \mathcal{O}\left(\frac{1}{(T+1)^{1/2}}\right) + \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}.$$

Summary of Sample Complexities

	Vanilla-PG	N-PG-IGT	(N)-HARPG
FOSP $\mathbb{E}[\|\nabla J(\theta)\|] \le \varepsilon$	$\tilde{\mathcal{O}}(\varepsilon^{-4})$ [1]	$\tilde{\mathcal{O}}(\varepsilon^{-3.5})$ (new)	$\tilde{\mathcal{O}}(\varepsilon^{-3})$ [5]
Global[(a)] $\mathbb{E}[J^* - J(\theta)] \le \varepsilon$	$\tilde{\mathcal{O}}(\varepsilon^{-3})$ [1]	$\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ (new)	$\tilde{\mathcal{O}}(\varepsilon^{-2})$ (new)

(a) Under Assumptions 1, 2; up to an error bar controlled by $\varepsilon_{\text{bias}}$.

Proof Sketch for N-PG-IGT

$$\text{Define:} \quad J_H(\theta) := \mathbb{E}_{\rho, \pi_\theta}\left[\sum_{h=0}^{H-1} \gamma^t r(s_h, a_h)\right].$$

Step I. Ascent-like lemma. If $\theta_{t+1} = \theta_t + \gamma_t \frac{d_t}{\|d_t\|}$, then

$$J(\theta_{t+1}) \ge J(\theta_t) + \frac{\gamma_t}{3}\|\nabla J(\theta_t)\| - \frac{8\gamma_t}{3}\|\hat{e}_t\| - \mathcal{O}(\gamma_t^2 + \gamma^H \gamma_t),$$

where $\hat{e}_t := d_t - \nabla J_H(\theta_t)$.

Step II. Relaxed weak gradient domination [3].

Lemma 1. Let Assumptions 1, 2 hold, and, in addition, $\|\nabla \log \pi_\theta(a|s)\| \le M_g$ for all $a \in \mathcal{A}$, $s \in \mathcal{S}$. Then

$$\varepsilon' + \|\nabla J(\theta)\| \ge \sqrt{2\mu}(J^* - J(\theta)) \quad \text{for all } \theta \in \mathbb{R}^d,$$

where $\varepsilon' := \frac{\mu_F\sqrt{\varepsilon_{\text{bias}}}}{M_g(1-\gamma)}$ and $\mu := \frac{\mu_F^2}{2M_g^2}$.

Step III. Variance reduction control.

$$\hat{e}_t = (1-\eta_t)\hat{e}_{t-1} + \eta_t e_t + (1-\eta_t)S_t + \eta_t Z_t,$$

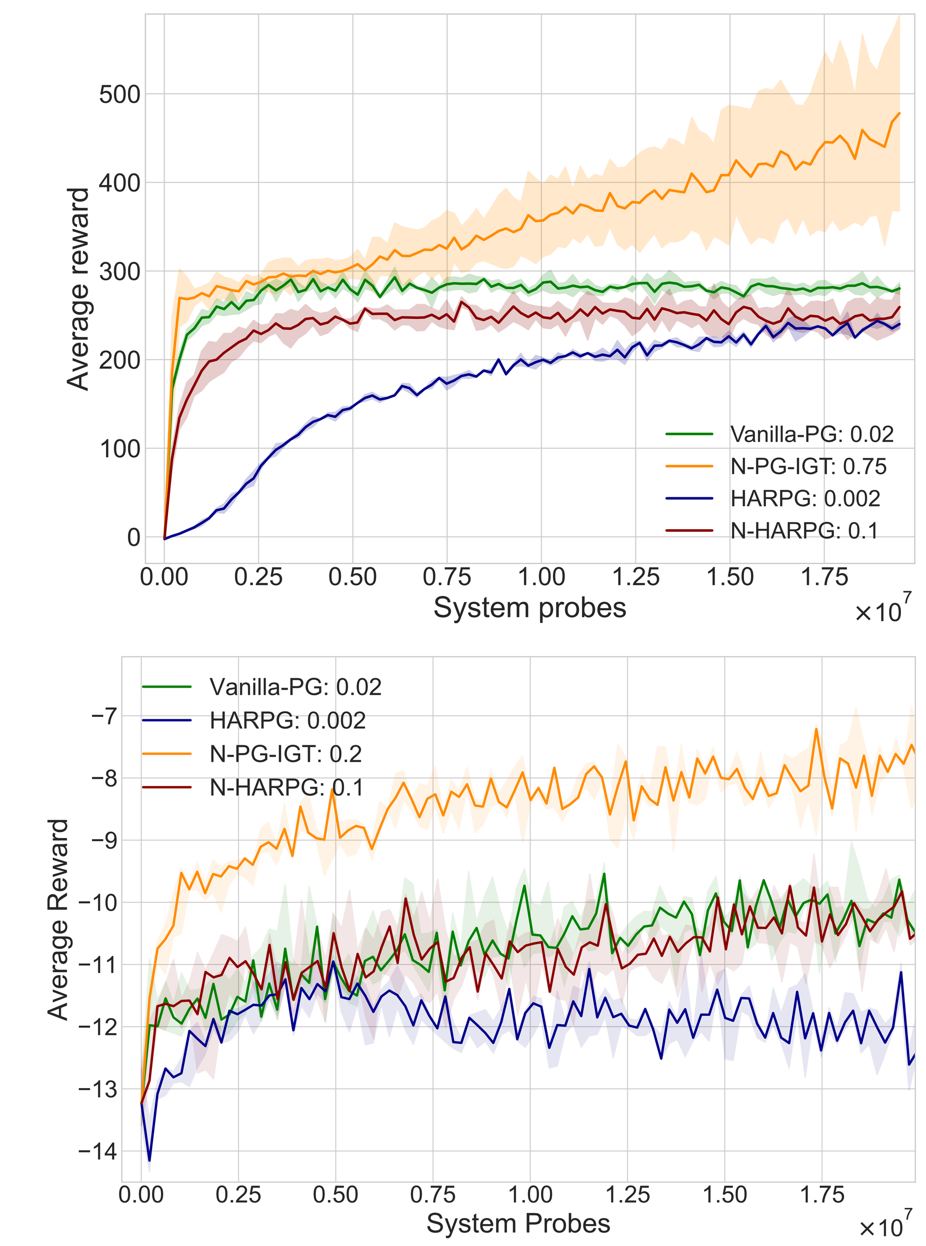
where $e_t := \widetilde{\nabla} J(\tilde{\tau}_t, \tilde{\theta}_t) - \nabla J_H(\tilde{\theta}_t)$, and S_t, Z_t are second-order Taylor approximation of J. $\mathbb{E}[\|\hat{e}_t\|] = \mathcal{O}\left(t^{-2/5}\right)$.

Step IV. Combine I-III and bound $\delta_t := \mathbb{E}[J^* - J(\theta_t)]$.

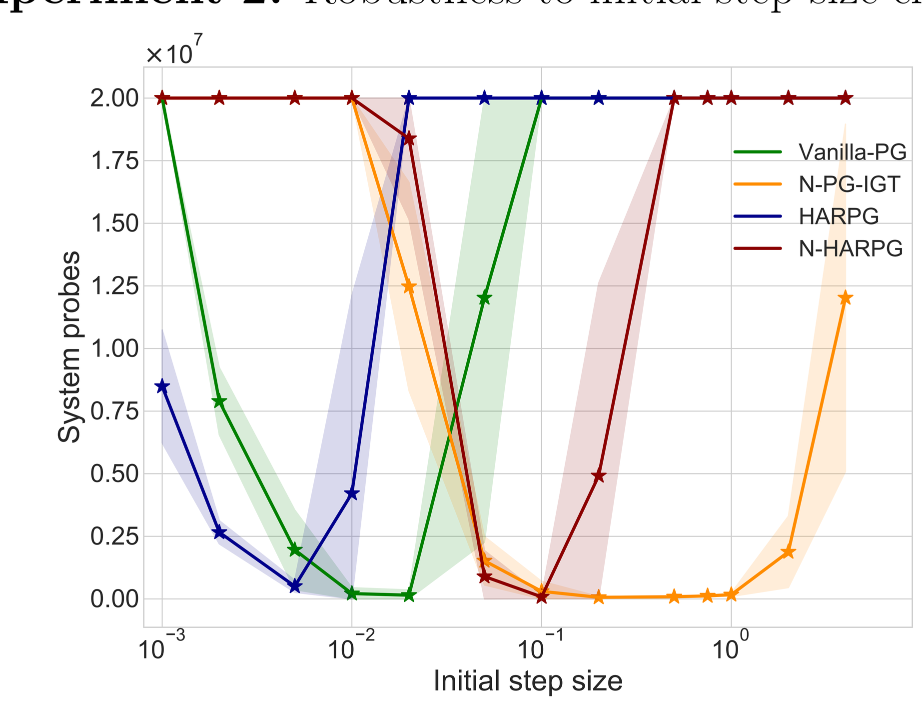
$$\delta_{t+1} \le (1 - \Omega(\gamma_t))\delta_t + \mathcal{O}(\gamma_t t^{-2/5} + \gamma_t^2 + \varepsilon'\gamma_t).$$

Experiments

Continuous control tasks: `Walker` (top), `Reacher`.
Policy Parameterization: $\pi_\theta(\cdot|s) \sim \mathcal{N}(\mu_\theta(s), \Sigma_\theta(s))$.
Experiment 1. Convergence with tuned initial γ_0.



Experiment 2. Robustness to initial step-size choice.



References

[1] R. Yuan, R. Gower, A. Lazaric. A general sample complexity analysis of vanilla policy gradient. AISTATS 2022.
[2] Y. Liu, K. Zhang, T. Basar, W. Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. NeurIPS 2020.
[3] Y. Ding, J. Zhang, J. Lavaei. On the global optimum convergence of momentum-based policy gradient. AISTATS 2022.
[4] A. Cutkosky, H. Mehta. Momentum improves normalized SGD. ICML 2020.
[5] S. Salehkaleybar, S. Khorasani, N. Kiyavash, N. He, P. Thiran. Momentum-Based Policy Gradient with Second-Order Information. arXiv:2205.08253, 2022.